

The hidden impact of inter-individual genomic variations on cellular function

Constantin Georgescu and Hamid Bolouri

Division of Biology, California Institute of Technology, CA 91125, USA

ABSTRACT

An analysis of the degree of genomic variation between two individual genomes suggests that there may be considerable biochemical differences among individuals. Examination of DNA sequence variations in 14 canonical signaling pathways and Monte-Carlo simulation modeling suggest that the kinetic and quantitative behavior of signaling pathways in many individuals may be significantly perturbed from the 'healthy' norm.

Signal transduction pathways in some individuals may suffer context-specific failures, or they may function normally but fail easily in the face of additional environmental perturbations or somatic mutations. These findings argue for new systems biology approaches that can predict pathway status in individuals using personal genome sequences and biomarker data.

Keywords: Genomics, systems biology, personal genomes, personalized medicine, genetics.

1. INTRODUCTION

Recent sequencing of multiple individual genomes has revealed enormous inter-individual differences. It is now estimated that the genomes of any two individuals will differ in as much as 0.5% of the genome¹.

The majority of an individual's sequence variants occur in genomic regions with no known function. Nonetheless, a large number of variants impact known regulatory and transcribed regions of genes. For example, the genome of Craig Venter¹ includes 6,114 non-synonymous SNPs (Single Nucleotide Polymorphisms) affecting the protein sequences of 4,107 genes (i.e. ~ 17% of the total).

Not all protein sequence changes will have a biochemical effect. However, additional sequence variations in 5' and 3' untranslated regions (UTRs), spliceosome recognition sites, and *cis*-regulatory regions* can also affect the behavior and function of gene products². All together, DNA sequence variations can lead to changes in binding affinity, gain or loss of molecular interaction partners, changes in half-life, and changes in spatial localization of the affected gene products.

Regulatory regions are thought to occupy a fraction of the genome comparable to the protein coding sequence². Therefore, variations in regulatory sequences may be expected to impact a similar proportion of all genes.

To what extent do such large-scale variations in individual genomes affect the disease susceptibilities and drug-responses of individuals? For Mendelian diseases (mostly single-gene disorders with simple patterns of inheritance), Venter's genome provides a hint: of the 4,107 genes with non-synonymous SNPs, 314 are associated with heritable traits in the OMIM database (<http://www.ncbi.nlm.nih.gov/omim/>) and 14 OMIM-

* Proximal promoters and distal enhancers typically span several hundred nucleotides and include dozens of DNA binding sites for transcriptional activators and repressors.

listed genes are impacted by copy number variations (CNVs comprise deletions and replications of DNA segments). The health implications of these variants can be predicted from the associated traits. To date, the health impacts of all other variants have been intractable.

Most diseases currently prevalent in industrialized countries (e.g. cardiovascular disease, neurodegenerative disease, diabetes) have multiple environmental and genetic components. Population studies suggest that susceptibility to and progression of many ‘multi-factorial’ diseases may arise from interactions among tens or even hundreds of genetic and environmental factors².

Responses to drugs also fall into single-gene (simple) and multi-gene (complex) categories. Many single-gene variations causing differences in drug uptake or clearance rates have been discovered². Genetic tests to identify different categories of responders, and drugs targeted to sub-populations carrying specific single-gene variants are becoming increasingly common (reviewed in reference 2). However, responses to drugs and other medical interventions can also vary due to complex interactions among multiple genes and environmental/life-style factors.

In the next section, we use the publicly-available personal genomes of James Watson and Craig Venter to quantify the extent to which inter-individual genomic differences impact specific cellular pathways. In section 3, we exploit the observation that cellular pathways are typically constructed from a collection of interacting functional building blocks³ to estimate the extent and type of impact genetic variations may have on the behavior of molecular pathways. Finally, section 4 explores the implications of our observations for biomedical research. In particular, we note that because of combinatorial differences in genomic variants, individuals exhibit considerable differences in susceptibilities and drug responses. We therefore end this paper with a call for new systems biology approaches combining personal genomics and personalized medicine.

2. THE POTENTIAL IMPACT OF GENOME VARIATIONS

Cellular state changes are typically mediated by inter-cellular signaling events. One way to evaluate the extent to which inter-individual genomic sequence variations may impact cellular function is to consider the potential effects of sequence variants on inter-cellular signaling.

The Kyoto Encyclopedia of Genes and Genomes (KEGG) currently has molecular-interaction network maps and related data for 14 canonical signaling pathways in humans. Many of these signaling systems (e.g. Wnt, Notch, Hedgehog) are used repeatedly in different cells and tissues of the body to effect a wide variety of cellular functions. Others such as the B-cell, T-cell and Toll-like receptor pathways are used to regulate specific and vital cellular functions such as adaptive and innate immunity.

Table 1 summarizes the extent of (potentially functionally significant) transcribed-sequence variations in 14 canonical signaling pathways for the individual genomes of James Watson⁴ and Craig Venter¹.

Variations in the introns and flanking (upstream and downstream) regions of genes – which may impact transcriptional regulation – are not listed in Table 1 because at present *cis*-regulatory sequences for most genes are poorly characterized. Since regulatory sequences are thought to occupy roughly the same length of DNA as coding regions², Table 1 underestimates the total amount of sequence variation affecting these signaling pathways by up to a factor of two.

We note from Table 1 that dozens to hundreds of transcribed-sequence variants impact each of the canonical signaling pathways in each individual. If we suppose that 1-in-10 of the variants listed in Table 1 has a biochemical effect on the relevant gene products, then every one of the pathways is likely to harbor multiple changes to its archetypal biochemical behavior.

Multiple variant proteins impacting a single signaling system could catastrophically affect its function. However, a number of mechanisms guard cellular physiology against catastrophic failures.

Table 1. The extent of DNA sequence variations in cellular signaling genes in two individuals (W=Watson, V=Venter). Abbreviations: BCR=B-cell receptor, Ca=Calcium, ErbB=the v-erb-b2 erythroblastic leukemia viral oncogene family (Receptor Tyrosine Kinases), Hh=Hedgehog, PI=Phosphatidylinositol, TCR=T-cell receptor, TLR=Toll-like receptor.

	frameshiftW	frameshiftV	nonsynonymousW	nonsynonymousV	UTR_W	UTR_V	splice_siteW	splice_siteV	No. genes
BCR	0	0	55	52	120	117	2	2	75
Ca	3	3	99	96	327	325	1	1	178
ErbB	3	3	29	24	161	187	0	1	87
Hh	0	0	24	32	74	83	1	3	56
JakStat	0	0	77	47	228	228	0	2	155
MAPK	1	0	128	127	436	451	9	3	269
mTOR	0	0	20	18	94	110	0	4	52
Notch	0	0	20	19	56	61	1	1	47
PI	0	0	55	61	159	175	0	0	76
TCR	0	0	49	42	153	163	1	0	108
TGFβ	0	0	27	28	85	139	0	1	87
TLR	0	0	43	39	126	139	12	8	101
VEGF	0	0	50	47	114	117	1	0	76
Wnt	0	2	55	51	279	246	1	2	163

The total number of genes in each canonical pathway is indicated in the right-most column of Table 1. Many of the genes assigned to each pathway have multiple isoforms. Normally only one of the isoforms is used in any particular cell type and condition. For example, there are 19 different genes encoding Wnt ligands in humans (see <http://tinyurl.com/WntHomePage> for details). Different members of the Wnt family are used to signal different types of events in different cells and conditions. Thus, the large numbers of variants per canonical pathway listed above translates to a smaller number of variants per signaling event in a particular cell. Moreover, other family members can often substitute for the loss of function of one their isoforms.

Thus, although several biochemically significant protein sequence variations may impact any given canonical signaling pathway, individual signaling events will be impacted by fewer variants. Moreover, any adverse effects caused by the variant may be diminished by the compensatory action of family member proteins (a form of graceful degradation mediated by redundancy).

The above considerations suggest that in general the enormous inter-individual variability in transcribed-region sequences does not lead to catastrophic health implications. On the other hand, the same considerations indicate that such sequence variations may have considerable kinetic and quantitative effects on particular cellular functions. In the next two sections, we expand on the latter finding and explore its implications in more detail.

3. COMBINATORIAL EFFECTS OF GENOMIC VARIATIONS

Figure 1, summarizes the extent of protein coding differences (frameshifts and non-synonymous SNPs) in each of the above signaling pathways for the Watson and Venter genomes (see figure legend for details).

In 13 of 14 pathways analyzed Venter's genome shares only one third of its variants with Watson's[†]. Thus, the particular combination of variants that may affect any specific signaling event in Venter is likely to be different from Watson's. Given that Watson and Venter are both of European ancestry, the level of person-to-person differences observed here may be representative for people from shared genetic backgrounds.

[†] Venter's diploid genome was sequenced to greater stringency than Watson's and therefore provides a better point of reference for our discussions here.

In support of the above observation, we note that the numbers of variants per individual genome is typically one or more orders of magnitude smaller than the total number of known commonly-occurring variants. Indeed, up to a quarter of variants in any individual genome may be low-frequency or rare, occurring in less than 5% and 1% of the population respectively².

Thus, the particular combination of sequence variants that govern cellular responses in a given individual are substantially different from person to person. Such person-specific combinatorial effects of sequence variants may explain some of the variability observed for a wide range of physiological parameters such as levels of blood C-reactive protein⁵ and cardiac Troponin⁶. Additional inter-individual differences arise from differences in exposures to environmental factors.

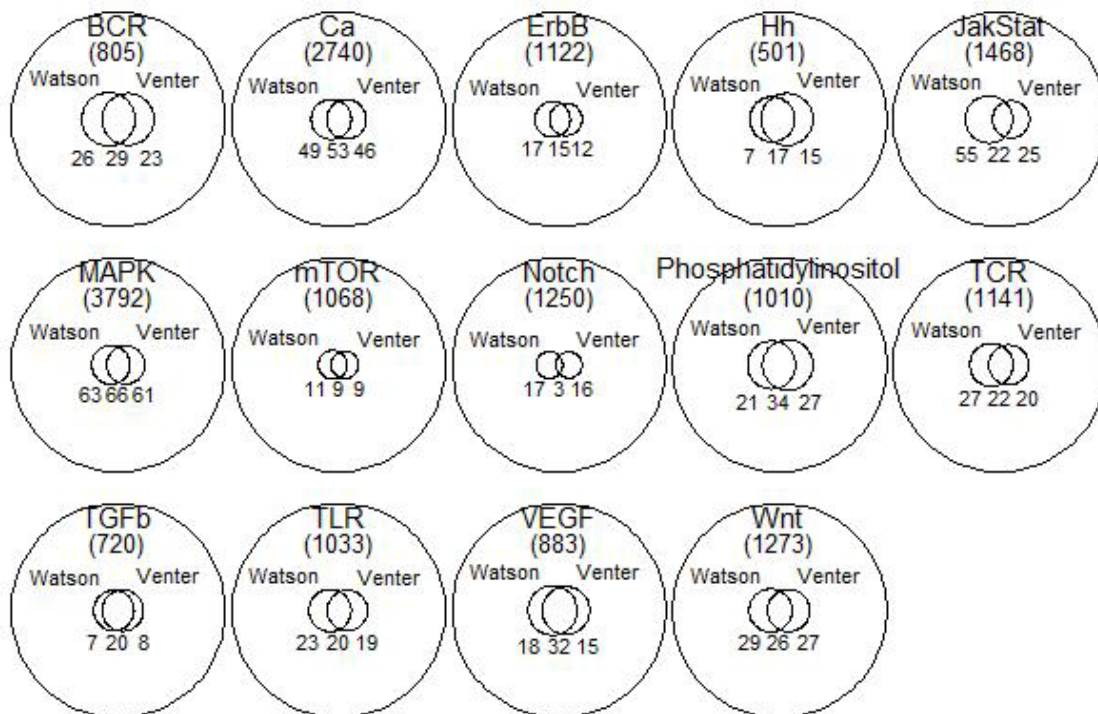


Figure 1. Number of DNA sequence variants (frameshifts and non-synonymous SNPs) affecting protein coding sequences of genes in 14 canonical signaling pathways. Pathway names are as indicated in Table 1. For each signaling pathway, the outer circle represents the full complement of variants known to affect genes in the respective pathway (corresponding to entries in the dbSNP database: <http://www.ncbi.nlm.nih.gov/projects/SNP/>). Total numbers of variants are given in brackets. The two inner circles represent the numbers of variants found in the Watson and Venter genomes. Circle sizes are proportional to the outer circle to provide a quick visual summary. The overlapping areas indicate shared variants between the two genomes. Numbers of variants not shared by the two genomes are indicated by the non-overlapping disk segments. Numbers of variants for each case are indicated below the corresponding segment. The Watson and Venter genomes include additional uncharacterized novel variants not shown here. All data were extracted from the ENSEMBL Human Variation 56 database using the ENSEMBL Biomart utility (<http://www.ensembl.org/>) via the R/Bioconductor biomaRt package (<http://tinyurl.com/biomaRt>).

Exposures to chemicals and other environmental factors can affect gene expression dramatically² in a dose-dependent manner. For example, high doses of Acetaminophen (the active ingredient in the commonly used painkiller Tylenol) affect the expression levels of 53 (i.e. about one third) of the genes in the Wnt signaling pathway (data from the Comparative Toxicogenomics Database, <http://ctd.mdibl.org/>). Indeed, TCD lists 2,896 substance-gene interactions affecting 146 out of 163 Wnt-pathway genes. The expression level of one gene common to Wnt and several other pathways (AKT1) is reportedly affected by 395 substances.

Dose-dependent gene expression changes resulting from exposures to many substances occur at all concentrations. Many environmental exposures (for example those arising from smoking, diet, and substance abuse) can have long-lasting effects. Thus, while low-level exposures have small effects and do not directly result in diseases, they confer life-history dependent differences in gene expression to individuals. In the next section, we argue that such differences in gene expression may contribute to differences in susceptibility to diseases.

In addition to the generic effects of environmental exposures on gene expression, particular genetic variants can confer greater or lesser susceptibility to some environmental insults. For example, smokers who carry the RS1051730 SNP are more susceptible to lung cancer than other smokers (data from SNPedia, <http://www.snpedia.com/>). Thus, an individual's particular combination of sequence variants can interact with their particular history of environmental exposures to result in highly individualized changes in gene expression for large numbers of genes.

In summary, environmental exposures can confer two levels of individuality in susceptibilities: Firstly, the history of each individual's exposures (and their effects) will be different for each individual. Secondly, genomic variations among individuals may lead to different responses to the same environmental factor.

4. THE POTENTIAL FOR HIDDEN SUSCEPTIBILITIES

Over the past decade, systems biology studies modeling the behaviors of multiple canonical signaling pathways^{7,8,9} have highlighted two commonly occurring types of response to external signals: bistability, and adaptation. In this section, we use simulation models of bistability and adaptation to explore the potential impact of combinatorial genetic and environmental perturbations on cellular function in individuals.

Bistability typically arises from positive feedback. Bistability in cellular pathways has been widely studied, and so will not be described in detail here (see references above and reference 3 for examples and detailed analysis). Consider the following minimal model of bistable response:

$$S' = \left(k_1 \cdot I + \frac{k_s \cdot S^H}{K_{MS} + S^H} \right) - k_{ds} \cdot S \quad (1)$$

Here, the variable I represents an input signal (e.g. the concentration of a ligand) and S represents the response of the signal transduction pathway to that input. k_1 , k_s , k_{ds} and K_{MS} represent kinetic parameters each aggregating multiple biochemical reaction rate constants (e.g. binding affinities and half lives). Figure 2 shows representative steady-state simulation results for a population of 250,000 such models.

To mimic the effects of multiple small genetic and environmental perturbations, we simultaneously varied all model parameters by a random amount uniformly distributed in the interval -0.1 to +0.1 (i.e. within $\pm 10\%$ of nominal values)^{*}.

The discontinuity in the output steady-states arises from the bistable behavior of the model. If we start with the system near the origin (low input and output levels) and increase the input level in a graded manner, the output level initially follows the lower curve until we reach the end of curve. Beyond this point (the low-to-high activation threshold, T_1) the steady state level of the output switches to the upper curve (the output high state). If we start with the system near the top right of the figure (input and output both high) and

^{*} The magnitude of variability selected here is not important to our conclusions. It is simply an arbitrary low-level of parameter variation to illustrate the principle that the co-occurrence of large numbers of small perturbations can lead to context-specific and hidden system susceptibilities.

gradually reduce the input signal level, the output will switch to the lower curve (low state) at the point where the upper curve ends (the high-to-low switching threshold, T_2).

A key characteristic of bistable switches is that $T_1 > T_2$, thus providing protection against noise in the input signal level. The difference between the output low and output high states provides another useful mechanism for system robustness: a well-defined noise margin between output low and output high levels (typically output high \gg output low).

All of the simulations in Figure 2 retain the intended bistable switching characteristic of the model. Note however, that two of the model instances (marked X) cannot be reset once they have been set to high (resetting would require $I < 0$, which is unphysiological, see figure legend for more details). The model instance marked ~ exhibits a more complex phenotype in that it can be reset only with an input level much below the nominal physiological low input level (left dashed vertical line).

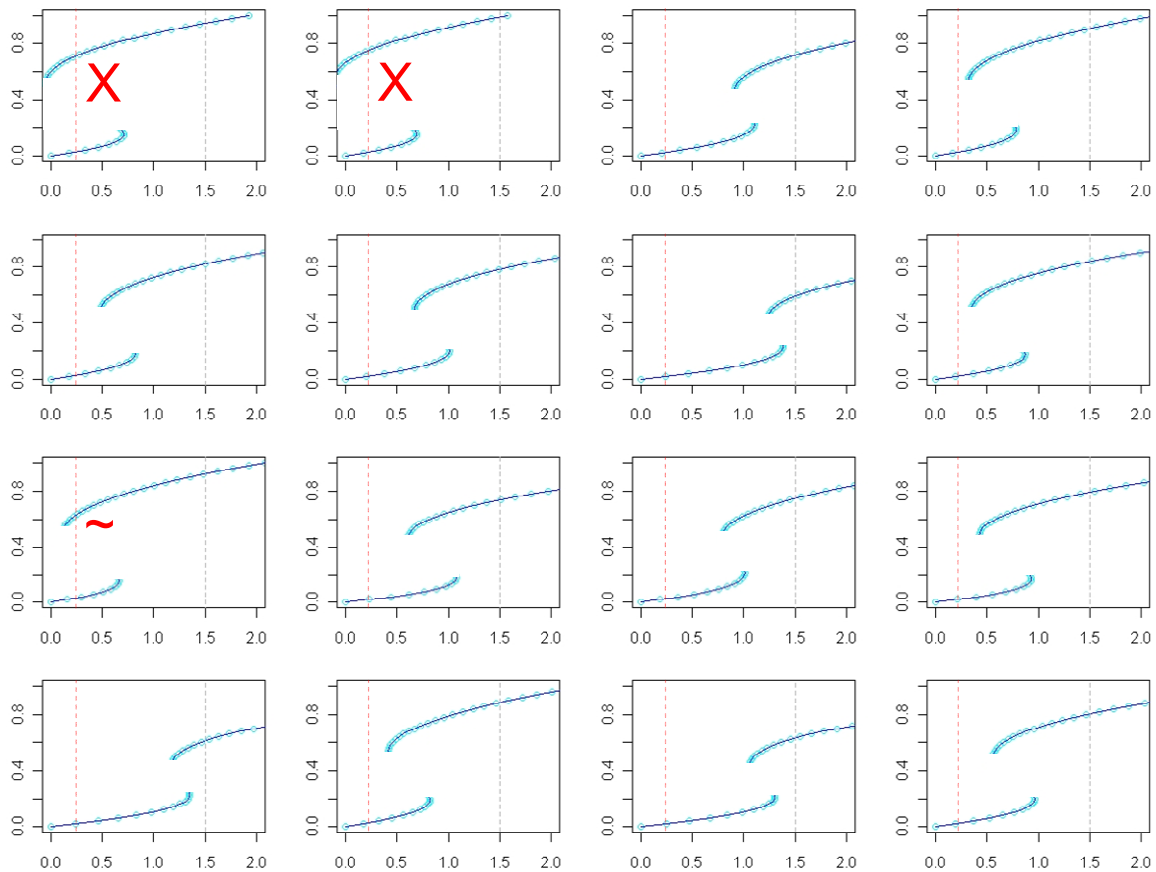


Figure 2. Representative simulation results from a population of 250,000 bistable-response signal transduction models. The horizontal axis in each plot represents the input signal level (arbitrary units). The vertical axis represents the output of the signal transduction pathway (arbitrary units). The dashed vertical lines represent nominal 'low' and 'high' input levels. The corresponding low and high output loci are represented by crossing points of the dashed lines with the gently rising output curves. The transitions of the output from low to high (by increasing the input signal level from nominal low to nominal high) and high to low (by reducing the input signal level) occur at different thresholds, as indicated by the overlap between the high-output and low-output curves. Note that for three of the model instances (marked X, X and ~), a low input level does not guarantee a low level of output activity (the dashed line representing a low input level crosses both loci). For these 'individuals' the signal transduction system will remain active even after the input signal has returned to nominally low levels (a gain of function situation similar to Wnt signaling in many

cancers). The system marked ~ can in fact be turned off by forcing the input (e.g. with drugs) to levels close to zero, i.e. well below the nominal 'low' signal. Once activated, the two systems marked X cannot be turned off at all.

The three model instances discussed above are illustrative of signaling pathways which are functional in some contexts and dysfunctional in other contexts. These systems will exhibit dysregulated behaviors under very specific conditions. All other model instances in figure 2 produce behaviors which are fully functional according to our characterization of bistable switch behavior (see preceding paragraph). However, several of these model instances (e.g. the top-right panel) have characteristics which may be considered close to failure. Such model instances may be considered 'predisposed' to failure. The corresponding signaling pathways could become dysregulated if environmental exposures or somatic mutations cause further perturbations to their operational parameters.

The histograms in figure 3 show the phenotype distributions for all 250,000 simulation instances. In all cases, the distributions are broad and long-tailed, indicating large differences between individuals and the presence of individuals with extreme phenotypes.

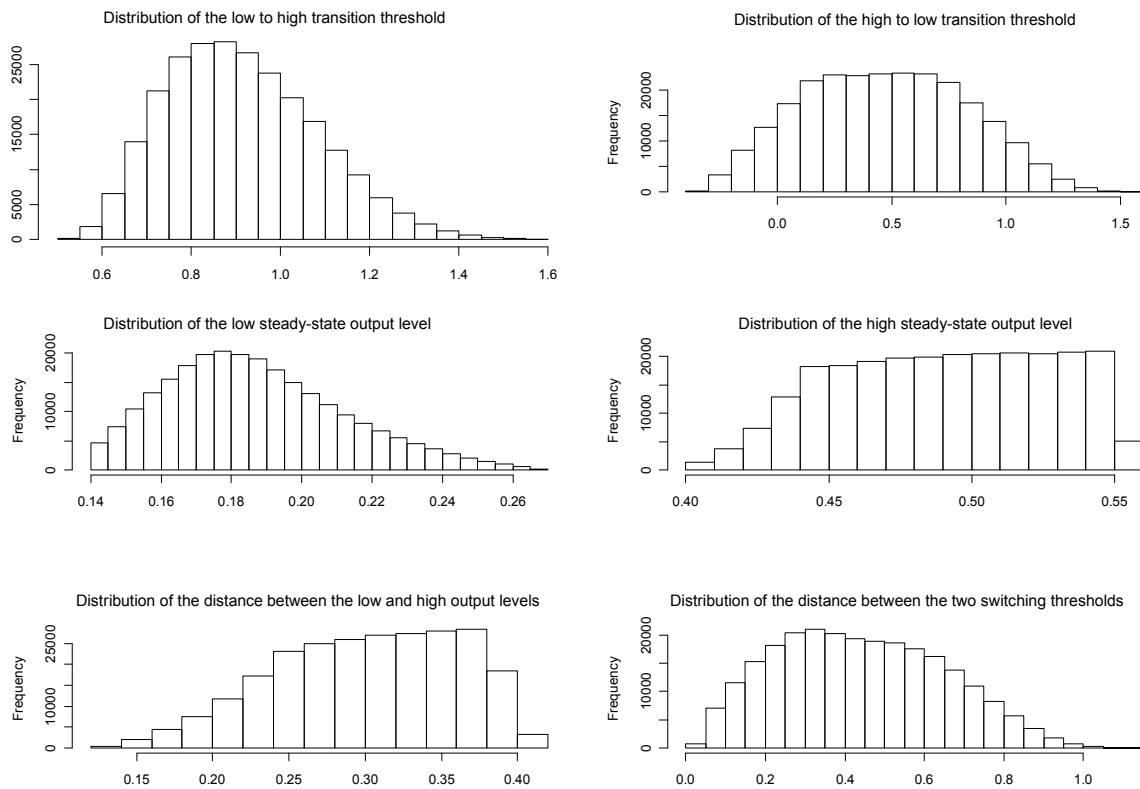


Figure 3, population characteristics of 250,000 simulated bistable signal transduction pathways. Each plot shows the histogram distribution for a different measure of bistability. In the histogram for the high-to-low transition threshold (T_2 , top right panel), a small fraction of the systems have $T_2 < 0$. These systems cannot be reset under physiological conditions, as discussed in Figure 2 and associated text. All distributions are broad and long-tailed.

To explore and illustrate the generality of the above conclusions, we performed a similar analysis for a minimal model of a signal transduction pathway with an adaptive response. Adaptive response models have been analyzed extensively elsewhere (see for example reference 10) and will therefore not be described in detail here. Our particular model is based on a network of interactions previously discovered downstream of TLR signaling in macrophages¹¹. The model is represented by the following equations:

$$\frac{dx}{dt} = \frac{k_1 \cdot \text{in}}{1 + k_i \cdot \text{in}} - k_2 \cdot x$$

$$\frac{dy}{dt} = \frac{k_3 \cdot \text{in}}{1 + k_i \cdot \text{in} + k_x \cdot x} - k_4 \cdot y \quad (2)$$

$$k_1, k_2 < k_3, k_4$$

The variable 'in' represents the input signal activity (e.g. ligand concentration). The inequality relation indicates that variable x responds more slowly than variable y. As with our previous model, the parameters (k_n) represent aggregate rate constants for multiple biochemical reactions.

Figure 4 shows representative simulation results for a population of such models. As before, all parameters have been varied randomly within $\pm 10\%$ of nominal. The input to the system (in, not shown in the plots) was activated at time zero and kept at a constant level throughout the simulations.

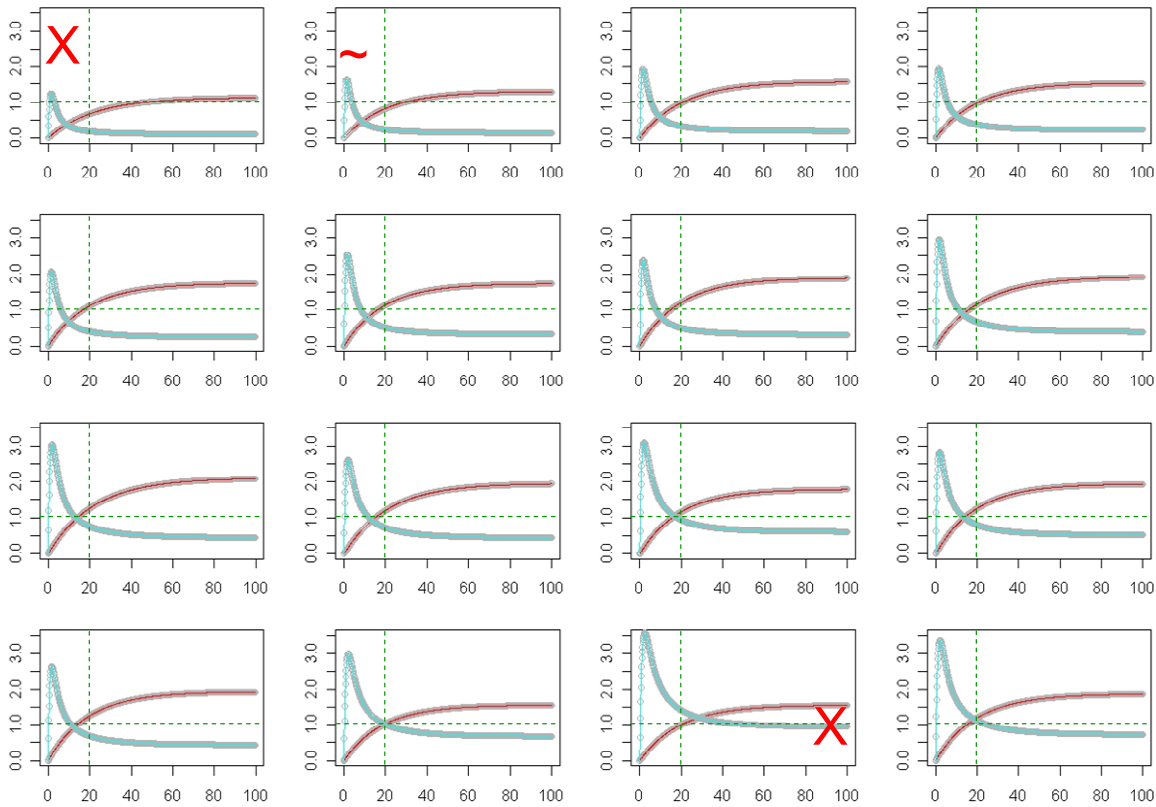


Figure 4. Representative simulation results for a population of adaptive signal transduction pathways. All parameters were varied randomly within $\pm 10\%$ of nominal. The input signal (not shown) is a step response. The horizontal axis represents time since the initiation of input activity (arbitrary units). The gently saturating curve shows the time course of the x variable (arbitrary units). The adaptive response of the y variable (peaked curve, arbitrary units) is evident in its sharp initial response and subsequent slow decline to near basal levels. The dashed lines mark nominal time and response magnitude checkpoints, as discussed in the text.

The dashed horizontal and vertical lines in Figure 4 mark arbitrarily selected time and response-magnitude values which we use to compare the performances of the various model instances. Suppose a ‘healthy’ response to a step input activated at time zero requires that the peak of y should exceed 2 units of magnitude, and that y should return to a magnitude less than 1 by time 20. Such requirements pose a minimal set of operational requirements on an adaptive response system.

For the simulation instances in Figure 4, we see that two model instances (marked X) fail these functional requirements. A third model instance (marked ~) produces a borderline response peak. Several of the other model runs are borderline in terms of time taken to relax back to basal activity levels (see for example the bottom-right panel). These model instances are analogous to apparently healthy cellular pathways with high sensitivity to new somatic mutations or environments insults.

The distributions of the operational characteristics of adaptive response models are shown in figure 5. Once again, the distributions are extremely broad (note the logarithmic frequency scale) and asymmetric, suggesting many individuals in the population are significantly different from the average.

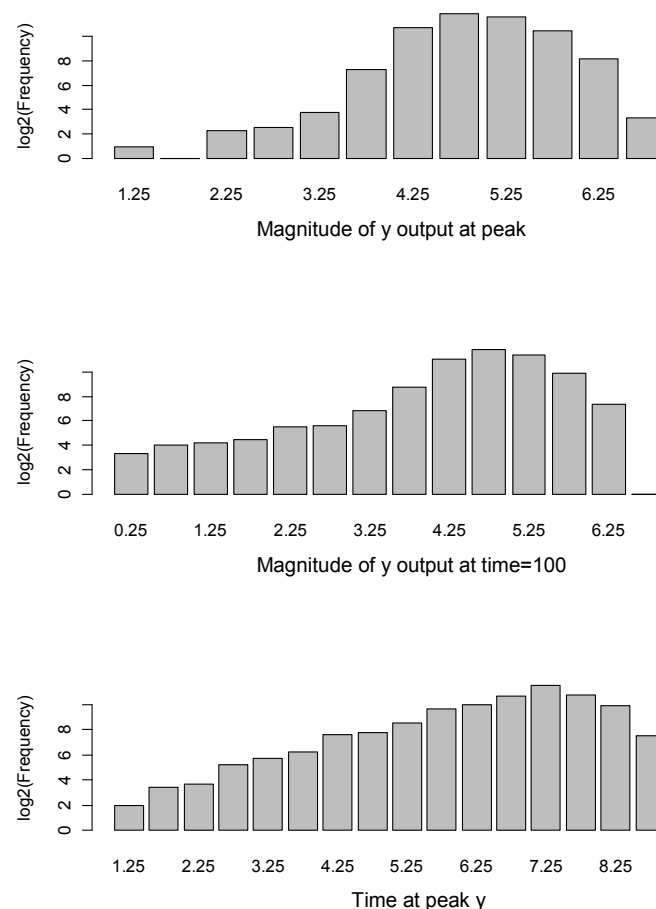


Figure 5. Variation of adaptation characteristics for 10,000 instances of the adaptive response model of Equation (2). The frequency (vertical) axis is plotted on a \log_2 scale in order to make the long tails of the distribution visible.

The above model simulations highlight two important insights into inter-individual differences in diseases susceptibility and drug response. Firstly, they illustrate how genomic variations and differences in environmental exposures can impact the function of cellular pathways and lead to different health outcomes

in individuals. Secondly, we note that not all healthy individuals are equal in terms of the levels and types of exposures that they can tolerate. Specifically, particular combinations of genetic and environmental perturbations can make some individuals more susceptible to health failures under additional stresses.

5. IMPLICATIONS FOR COMPUTATIONAL BIOLOGY RESEARCH

We have argued that combinatorial differences in DNA sequence variations lead to large and widespread inter-individual differences in the kinetic and quantitative behavior of cellular pathways. The general nature of our Monte-Carlo simulations of bistable and adaptive signaling systems suggests that the predisposing and hidden effects of DNA sequence variations and environmental exposures are likely to impact many cellular pathways and be widespread in populations.

Our analyses of the available data suggest that each individual's *combination* of genetic variations and life history exposures is unique. Diagnostic and treatment approaches that ignore inter-individual differences in susceptibility and treatment-response may be convenient generalizations that are far from optimal for individual patients.

It appears that susceptibility differences may play a much larger role in determining health status than has so far been recognized. This observation implies an urgent need to detect key susceptibilities in affected individuals. It is therefore incumbent on biomedical research to extend systems biology approaches to characterizing pathways in individuals. Such approaches will require the characterization not only of personal genomes, but also of large amounts of biomarker data that must be collected in order to quantify the impact of past environmental exposures.

There is a pressing need for medicine to transition from its current mode of reacting to symptoms long-after the onset of diseases, to a pro-active approach that identifies the *onset* of diseases at the earliest opportunity. To detect diseases before the onset of obvious symptoms, it will be necessary to characterize and monitor individual susceptibilities.

The technology to sequence individual genomes is already here and will reach cost levels that will allow its widespread use in the next few years². Requisite minimally-invasive, reliable and low-cost biomonitoring technologies such as ultra-sound imaging, optical imaging, and blood proteome analysis are currently under development and show great promise². Data from these technologies tends to be large in volume and highly complex in nature. Sophisticated new algorithms will be needed in order to process biomarker data for large numbers of patients and perform classification tests with acceptably low false alarm rates and manpower and time overheads.

Systems biology is a relatively young discipline. Its application to medicine will impose heavy demands on both disciplines and drive systems biology approaches to a more sophisticated maturity.

Acknowledgement: This work was supported in part by NHLBI grant number HL089102 to HB.

REFERENCES

- [1] Levy S et al, The Diploid Genome Sequence of an Individual Human, PLoS Biology, 5(10):e254, (2007).
- [2] Bolouri H, Personal Genomics and Personalized Medicine, Imperial College Press, London, 2009.
- [3] Bolouri H, Computational Modeling of Gene Regulatory Networks – A Primer, Imperial College Press, London, 2008.
- [4] Wheeler DA et al, The complete genome of an individual by massively parallel DNA sequencing, Nature, (452):872-877, (2008).

- [5] Macy EM, Hayes TE, and Tracy RP, Variability in the measurement of C-reactive protein in healthy subjects: implications for reference intervals and epidemiological applications, *Clinical Chemistry*, 43(1):52–58, (1997).
- [6] Antman EM et al, Cardiac-specific troponin I levels to predict the risk of mortality in patients with acute coronary syndromes, *New England Journal of Medicine*, (335):1342-1349, (1996).
- [7] Schoeberl B, Eichler-Jonsson C, Gilles ED, Müller G, Computational modeling of the dynamics of the MAP kinase cascade activated by surface and internalized EGF receptors, *Nature Biotechnology*, 20(4):370-375, 2002.
- [8] Bhalla US, Ram PT, Iyengar R, MAP kinase phosphatase as a locus of flexibility in a mitogen-activated protein kinase signaling network, *Science*, 297(5583):1018-23, (2002).
- [9] Vilar JMG, Jansen R, Sander C, Signal processing in the TGF-beta superfamily ligand-receptor network, *PLoS Computational Biology*, 2(1):e3, (2006).
- [10] Ma W et al, Defining network topologies that can achieve biochemical adaptation, *Cell*, 138(4):760-773, (2009).
- [11] Gilchrist M, et al, Systems biology approaches identify ATF3 as a negative regulator of innate immunity, *Nature*, 441(7090):173-178, (2006).